



## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>7</sup> :</b> <b>C12N 15/12, C07K 14/47, 16/18, G01N 33/566, C12Q 1/68, C12N 15/11, 15/62, A01K 67/027, A61K 38/00</b>	<b>A2</b>	<b>(11) International Publication Number:</b> <b>WO 00/58473</b> <b>(43) International Publication Date:</b> 5 October 2000 (05.10.00)																																				
<b>(21) International Application Number:</b> PCT/US00/08621 <b>(22) International Filing Date:</b> 31 March 2000 (31.03.00)  <b>(30) Priority Data:</b> <table border="0" style="width: 100%;"> <tr> <td style="width: 30%;">60/127,607</td> <td style="width: 40%;">31 March 1999 (31.03.99)</td> <td style="width: 30%;">US</td> </tr> <tr> <td>60/127,636</td> <td>2 April 1999 (02.04.99)</td> <td>US</td> </tr> <tr> <td>60/127,728</td> <td>5 April 1999 (05.04.99)</td> <td>US</td> </tr> <tr> <td>09/540,763</td> <td>30 March 2000 (30.03.00)</td> <td>US</td> </tr> </table> <b>(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Applications</b> <table border="0" style="width: 100%;"> <tr> <td style="width: 30%;">US</td> <td style="width: 40%;">60/127,607 (CIP)</td> <td style="width: 30%;"></td> </tr> <tr> <td>Filed on</td> <td>31 March 1999 (31.03.99)</td> <td></td> </tr> <tr> <td>US</td> <td>60/127,636 (CIP)</td> <td></td> </tr> <tr> <td>Filed on</td> <td>2 April 1999 (02.04.99)</td> <td></td> </tr> <tr> <td>US</td> <td>60/127,728 (CIP)</td> <td></td> </tr> <tr> <td>Filed on</td> <td>5 April 1999 (05.04.99)</td> <td></td> </tr> <tr> <td>US</td> <td>09/540,763 (CIP)</td> <td></td> </tr> <tr> <td>Filed on</td> <td>30 March 2000 (30.03.00)</td> <td></td> </tr> </table> <b>(71) Applicant (for all designated States except US):</b> CURAGEN CORPORATION [US/US]; 555 Long Wharf Drive, 11th Floor, New Haven, CT 06511 (US).		60/127,607	31 March 1999 (31.03.99)	US	60/127,636	2 April 1999 (02.04.99)	US	60/127,728	5 April 1999 (05.04.99)	US	09/540,763	30 March 2000 (30.03.00)	US	US	60/127,607 (CIP)		Filed on	31 March 1999 (31.03.99)		US	60/127,636 (CIP)		Filed on	2 April 1999 (02.04.99)		US	60/127,728 (CIP)		Filed on	5 April 1999 (05.04.99)		US	09/540,763 (CIP)		Filed on	30 March 2000 (30.03.00)		<b>(72) Inventors; and</b> <b>(75) Inventors/Applicants (for US only):</b> SHIMKETS, Richard, A. [US/US]; 191 Leete Street, West Haven, CT 06516 (US). LEACH, Martin [GB/US]; 884 School Street, Webster, MA 01570 (US).  <b>(74) Agent:</b> ELRIFI, Ivor, R.; Mintz, Levin, Cohn, Ferris, Glovsky and Popeo, P.C., One Financial Center, Boston, MA 02111 (US).  <b>(81) Designated States:</b> AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).  <b>Published</b> <i>Without international search report and to be republished upon receipt of that report.</i>
60/127,607	31 March 1999 (31.03.99)	US																																				
60/127,636	2 April 1999 (02.04.99)	US																																				
60/127,728	5 April 1999 (05.04.99)	US																																				
09/540,763	30 March 2000 (30.03.00)	US																																				
US	60/127,607 (CIP)																																					
Filed on	31 March 1999 (31.03.99)																																					
US	60/127,636 (CIP)																																					
Filed on	2 April 1999 (02.04.99)																																					
US	60/127,728 (CIP)																																					
Filed on	5 April 1999 (05.04.99)																																					
US	09/540,763 (CIP)																																					
Filed on	30 March 2000 (30.03.00)																																					
<b>(54) Title:</b> NUCLEIC ACIDS INCLUDING OPEN READING FRAMES ENCODING POLYPEPTIDES; "ORFX"  <b>(57) Abstract</b> <p>The present invention provides open reading frames ORFX, encoding isolated polypeptides, as well as polynucleotides encoding ORFX and antibodies that immunospecifically bind to ORFX or any derivative, variant, mutant, or fragment of the ORFX polypeptides, polynucleotides or antibodies. The invention additionally provides methods in which the ORFX polypeptide, polynucleotide and antibody are used in detection and treatment of a broad range of pathological states, as well as to other uses.</p>																																						

## NOVEL POLYNUCLEOTIDES AND POLYPEPTIDES ENCODED THEREBY

5

## BACKGROUND OF THE INVENTION

The invention relates generally to nucleic acids and polypeptides encoded thereby, and methods of using these nucleic acids and polypeptides.

10

## SUMMARY OF THE INVENTION

The invention is based in part on the discovery of nucleic acids that include open reading frames encoding novel polypeptides, and on the polypeptides encoded thereby. The nucleic acids and polypeptides are collectively referred to herein as "ORFX".

15

Accordingly, in one aspect, the invention provides an isolated nucleic acid molecule (SEQ ID NO:2 $n$ -1, wherein  $n$  is an integer between 1-3161), that encodes novel polypeptide, or a fragment, homolog, analog or derivative thereof. The nucleic acid can include, *e.g.*, a nucleic acid sequence encoding a polypeptide at least 85% identical to a polypeptide comprising the amino acid sequences of SEQ ID NO:2 $n$ , wherein  $n$  is an integer between 1-3161. The nucleic acid can be, *e.g.*, a genomic DNA fragment, or a cDNA molecule.

20

Also included in the invention is a vector containing one or more of the nucleic acids described herein, and a cell containing the vectors or nucleic acids described herein.

The invention is also directed to host cells transformed with a recombinant expression vector comprising any of the nucleic acid molecules described above.

25

In another aspect, the invention includes a pharmaceutical composition that includes an ORFX nucleic acid and a pharmaceutically acceptable carrier or diluent.

In a further aspect, the invention includes a substantially purified ORF polypeptide, *e.g.*, any of the ORFX polypeptides encoded by an ORFX nucleic acid, and fragments, homologs, analogs, and derivatives thereof. The invention also includes a pharmaceutical composition that includes a ORFX polypeptide and a pharmaceutically acceptable carrier or diluent.

5 In a still a further aspect, the invention provides an antibody that binds specifically to an ORFX polypeptide. The antibody can be, *e.g.*, a monoclonal or polyclonal antibody, and fragments, homologs, analogs, and derivatives thereof. The invention also includes a pharmaceutical composition including ORFX antibody and a pharmaceutically acceptable carrier or diluent. The invention is also directed to isolated antibodies that bind to an epitope on a  
10 polypeptide encoded by any of the nucleic acid molecules described above.

The invention also includes kits comprising any of the pharmaceutical compositions described above.

The invention further provides a method for producing an ORFX polypeptide by providing a cell containing a ORFX nucleic acid, *e.g.*, a vector that includes a ORFX nucleic  
15 acid, and culturing the cell under conditions sufficient to express the ORFX polypeptide encoded by the nucleic acid. The expressed ORFX polypeptide is then recovered from the cell. Preferably, the cell produces little or no endogenous ORFX polypeptide. The cell can be, *e.g.*, a prokaryotic cell or eukaryotic cell.

The invention is also directed to methods of identifying an ORFX polypeptide or nucleic  
20 acids in a sample by contacting the sample with a compound that specifically binds to the polypeptide or nucleic acid, and detecting complex formation, if present.

The invention further provides methods of identifying a compound that modulates the activity of a ORFX polypeptide by contacting ORFX polypeptide with a compound and determining whether the ORFX polypeptide activity is modified.

25 The invention is also directed to compounds that modulate ORFX polypeptide activity identified by contacting a ORFX polypeptide with the compound and determining whether the compound modifies activity of the ORFX polypeptide, binds to the ORFX polypeptide, or binds to a nucleic acid molecule encoding a ORFX polypeptide.

In a another aspect, the invention provides a method of determining the presence of or  
30 predisposition of an ORFX-associated disorder in a subject. The method includes providing a sample from the subject and measuring the amount of ORFX polypeptide in the subject sample.

The amount of ORFX polypeptide in the subject sample is then compared to the amount of ORFX polypeptide in a control sample. An alteration in the amount of ORFX polypeptide in the subject protein sample relative to the amount of ORFX polypeptide in the control protein sample indicates the subject has a tissue proliferation-associated condition. A control sample is preferably taken from a matched individual, *i.e.*, an individual of similar age, sex, or other general condition but who is not suspected of having a tissue proliferation-associated condition. Alternatively, the control sample may be taken from the subject at a time when the subject is not suspected of having a tissue proliferation-associated disorder. In some embodiments, the ORFX is detected using a ORFX antibody.

In a further aspect, the invention provides a method of determining the presence of or predisposition of an ORFX-associated disorder in a subject. The method includes providing a nucleic acid sample, *e.g.*, RNA or DNA, or both, from the subject and measuring the amount of the ORFX nucleic acid in the subject nucleic acid sample. The amount of ORFX nucleic acid sample in the subject nucleic acid is then compared to the amount of an ORFX nucleic acid in a control sample. An alteration in the amount of ORFX nucleic acid in the sample relative to the amount of ORFX in the control sample indicates the subject has a tissue proliferation-associated disorder.

In a still further aspect, the invention provides method of treating or preventing or delaying a ORFX-associated disorder. The method includes administering to a subject in which such treatment or prevention or delay is desired a ORFX nucleic acid, a ORFX polypeptide, or an ORFX antibody in an amount sufficient to treat, prevent, or delay a tissue proliferation-associated disorder in the subject.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In the case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the following detailed description and claims.

## DETAILED DESCRIPTION OF THE INVENTION

The invention provides novel polypeptides and nucleotides encoded thereby. The polynucleotides and their encoded polypeptides can be grouped according to the functions played by their gene products. Such functions include, structural proteins, proteins from which associated with metabolic pathways fatty acid metabolism, glycolysis, intermediary metabolism, calcium metabolism, proteases, and amino acid metabolism, etc.

Included in the invention are 3161 novel nucleic acid sequences and their encoded polypeptides. The sequences are collectively referred to as "ORFX nucleic acids" or ORFX polynucleotides" and the corresponding encoded polypeptide is referred to as a "ORFX polypeptide" or ORFX protein". For example, an ORFX nucleic acid according to the invention is a nucleic acid including an ORF1 nucleic acid, and an ORF polypeptide according to the invention is a polypeptide that includes the amino acid sequence of an ORF1 polypeptide. Unless indicated otherwise, "ORFX" is meant to refer to any of the ORF1-3161 sequences disclosed herein.

Table 1 provides a summary of the ORFX nucleic acids and their encoded polypeptides are summarized in Table 1. Nucleic acid sequences and polypeptide sequences for ORFX nucleic acids according to the invention is provided in the section of the specification entitled "Disclosed Sequences of ORFX Nucleic Acid and Polypeptide Sequences."

Column 1 of Table 1, entitled "ORF #", denotes an ORF number assigned to a nucleic acid containing an open reading frame according to the invention.

Column 2 of Table 1, entitled "Internal Identification number (Nucleic Acid Sequence Identification Number, Polypeptide Sequence Identification Number), provides an internal identification number for the indicated ORF, along with sequence identification numbers (SEQ ID NOs.) corresponding to the indicated ORF. In general, for an ORF $n$  according to the invention (wherein  $n$  is any integer from 1 to 3161), a nucleic acid corresponding to the ORF is SEQ ID NO:2 $n$ -1, and an amino acid sequence encoded by the ORF is SEQ ID NO:2 $n$ . For example, a nucleic acid sequence corresponding to an ORF1 nucleic acid is SEQ ID NO:1, and a polypeptide sequence corresponding to an ORF1 polypeptide is SEQ ID NO:2. Similarly, a

nucleic acid sequence corresponding to an ORF4 nucleic acid is SEQ ID NO:7, and a polypeptide sequence corresponding to an ORF4 polypeptide is SEQ ID NO:8; a nucleic acid sequence corresponding to an ORF198 nucleic acid sequence is SEQ ID NO:395, and a polypeptide sequence corresponding to an ORF198 polypeptide is SEQ ID NO:396. Nucleic acid sequences and polypeptide sequences for ORFX nucleic acids according to the invention are provided in the section of the specification entitled "Disclosed Sequences of ORFX Nucleic Acid and Polypeptide Sequences."

Column 2 of Table 1, entitled "Protein Similarity", lists previously described proteins that are related to polypeptides encoded by the ORFs. Genbank identifiers for the previously described proteins are provided. These can be retrieved from <http://www.ncbi.nlm.nih.gov/>.

To determine similarity to previously described proteins, polypeptides encoded by ORFX DNA sequences were tested using the Framesearch Algorithm against a nonredundant version of the GenPept Database from NCBI/Genbank. DNA sequences that had a score of '90' or above (Framesearch algorithm score, Edelman et. al. GCG Genetics) to a known protein were selected. Open reading frames were extended beyond the region of the protein matched using standard DNA translation and codon tables. Novel proteins that lacked a protein match were translated against the standard genetic codons and proteins with an ORF at least 80 amino acids and containing a Methionine start are included in the Table.

Column 3 of Table 3, entitled "Protein Domains", lists previously described protein domains, designated by pfam entries, that are present in polypeptides encoded by the ORFs. Also included in column 3 are proteins in which these domains are present. The pfam entries can be retrieved from <http://pfam.wustl.edu/>. DNA sequences were translated in all six frames and tested using the Hmmer Algorithm against the Pfam Database (References to the algorithm and Pfam database can be found at <http://pfam.wustl.edu>). Translated DNA sequences that matched a protein domain entry in the Pfam database AND had a score of '7.5' were selected.

Column 4 of Table 3, entitled "Protein Classification", lists the type of classification assigned for the protein, based on its homology. Examples of proteins in the classification include the following proteins:

### **Amylases**

Amylase is responsible for endohydrolysis of 1,4-alpha-glucosidic linkages in oligosaccharides and polysaccharides. Variations in amylase gene may be indicative of delayed maturation and of various amylase producing neoplasms and carcinomas.

### **5 Amyloid**

The serum amyloid A (SAA) proteins comprise a family of vertebrate proteins that associate predominantly with high density lipoproteins (HDL). The synthesis of certain members of the family is greatly increased in inflammation. Prolonged elevation of plasma SAA levels, as in chronic inflammation, 15 results in a pathological condition, called amyloidosis, which affects the liver, kidney and spleen and which is characterized by the highly insoluble accumulation of SAA in these tissues. Amyloid selectively inhibits insulin-stimulated glucose utilization and glycogen deposition in muscle, while not affecting adipocyte glucose metabolism. Deposition of fibrillar amyloid proteins intraneuronally, as neurofibrillary tangles, extracellularly, as plaques and in blood vessels, is characteristic of both Alzheimer's disease and aged Down's syndrome. 15 Amyloid deposition is also associated with type II diabetes mellitus.

### **Angiopoeitin**

Members of the angiopoeitin/fibrinogen family have been shown to stimulate the generation of new blood vessels, inhibit the generation of new blood vessels, and perform several roles in blood clotting. This generation of new blood vessels, called angiogenesis, is also an 20 essential step in tumor growth in order for the tumor to get the blood supply it needs to expand. Variation in these genes may be predictive of any form of heart disease, numerous blood clotting disorders, stroke, hypertension and predisposition to tumor formation and metastasis. In particular, these variants may be predictive of the response to various antihypertensive drugs and chemotherapeutic and anti-tumor agents.

### **25 Apoptosis-related proteins**

Active cell suicide (apoptosis) is induced by events such as growth factor withdrawal and toxins. It is controlled by regulators, which have either an inhibitory effect on programmed cell

death (anti-apoptotic) or block the protective effect of inhibitors (pro-apoptotic). Many viruses have found a way of countering defensive apoptosis by encoding their own anti-apoptosis genes preventing their target-cells from dying too soon. Variants of apoptosis related genes may be useful in formulation of anti-aging drugs.

5           **Cadherin, Cyclin, Polymerase, Oncogenes, Histones, Kinases**

Members of the cell division/cell cycle pathways such as cyclins, many transcription factors and kinases, DNA polymerases, histones, helicases and other oncogenes play a critical role in carcinogenesis where the uncontrolled proliferation of cells leads to tumor formation and eventually metastasis. Variation in these genes may be predictive of predisposition to any form  
10 of cancer, from increased risk of tumor formation to increased rate of metastasis. In particular, these variants may be predictive of the response to various chemotherapeutic and anti-tumor agents.

**Colony-stimulating factor-related proteins**

Granulocyte/macrophage colony-stimulating factors are cytokines that act in  
15 hematopoiesis by controlling the production, differentiation, and function of 2 related white cell populations of the blood, the granulocytes and the monocytes-macrophages.

**Complement-related proteins**

Complement proteins are immune associated cytotoxic agents, acting in a chain reaction to exterminate target cells to that were opsonized (primed) with antibodies, by forming a  
20 membrane attack complex (MAC). The mechanism of killing is by opening pores in the target cell membrane. Variations in 20 complement genes or their inhibitors are associated with many autoimmune disorders. Modified serum levels of complement products cause edemas of various tissues, lupus (SLE), vasculitis, glomerulonephritis, renal failure, hemolytic anemia, thrombocytopenia, and arthritis. They interfere with mechanisms of ADCC (antibody dependent  
25 cell cytotoxicity), severely impair immune competence and reduce phagocytic ability. Variants of complement genes may also be indicative of type I diabetes mellitus, meningitis neurological disorders such as nemaline myopathy, neonatal hypotonia, muscular disorders such as congenital myopathy and other diseases.

### **Cytochrome**

The respiratory chain is a key biochemical pathway which is essential to all aerobic cells. There are five different cytochromes involved in the chain. These are heme bound proteins which serve as electron carriers. Modifications in these genes may be predictive of ataxia areflexia, dementia and myopathic and neuropathic changes in muscles. Also, association with various types of solid tumors.

### **Kinesins**

Kinesins are tubulin molecular motors that function to transport organelles within cells and to move chromosomes along microtubules during cell division. Modifications of these genes may be indicative of neurological disorders such as Pick disease of the brain, tuberous sclerosis.

### **Cytokines, Interferon, Interleukin**

Members of the cytokine families are known for their potent ability to stimulate cell growth and division even at low concentrations. Cytokines such as erythropoietin are cell-specific in their growth stimulation; erythropoietin is useful for the stimulation of the proliferation of erythroblasts. Variants in cytokines may be predictive for a wide variety of diseases, including cancer predisposition.

### **G-protein coupled receptors**

G-protein coupled receptors (also called R7G) are an extensive group of hormones, neurotransmitters, odorants and light receptors which transduce extracellular signals by interaction with guanine nucleotide-binding (G) proteins. Alterations in genes coding for G-coupled proteins may be involved in and indicative of a vast number of physiological conditions. These include blood pressure regulation, renal dysfunctions, male infertility, dopamine associated cognitive, emotional, and endocrine functions, hypercalcemia, chondrodysplasia and osteoporosis, pseudohypoparathyroidism, growth retardation and dwarfism.

**Thioesterases**

Eukaryotic thiol proteases are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. Variants of thioester associated genes may be predictive of neuronal disorders and mental illnesses such as Ceroid Lipoffiscinosis, Neuronal 1, Infantile, Santavuori disease and more.

The key to the molecule type is as follows:

10	Abbrev:	Title:
	amylase	amylase protein
	amylaseinhib	amylase inhibitor
	amyloid	amyloid protein
15	apoptosis	apoptosis associated protein
	apoptosisinhib	apoptosis inhibitors
	apoptosisrecep	apoptosis receptors
	ATPase_associated	ATPase associated protein
	biotindep	biotin dependent enzyme/protein
20	cadherin	cadherin protein
	calcium_channel	calcium channel protein
	carboxylase	carboxylase protein
	cathepsin	cathepsin/carboxypeptidases
	cathepsininhib	cathepsin/carboxypeptidase inhibitor
25	chloride_channel	chloride channel protein
	collagen	collagen
	complement	complement protein
	complementrecept	complement receptor protein
	complementinhib	complement inhibitor
30	csf	colony stimulating factor
	csfrecept	colony stimulating factor receptor
	cyclin	cyclin protein
	cyto450	cytochrome p450 protein
	cytochrome	cytochrome related protein
35	deaminase	deaminase
	dehydrogenase	dehydrogenase
	desaturase	desaturase
	dna_rna_bind	DNA/RNA binding protein/factor
	dna_rna_inhib	DNA/RNA binding protein/factor inhibitor
40	dynein	dynein

	elastase	elastase
	elastaseinhib	elastase inhibitor
	eph	EPH family of tyrosine kinases
	esterase	esterase
5	esteraseinhib	esterase inhibitor
	fgf	fibroblast growth factor
	fgfreceptor	fibroblast growth factor receptor
	gaba	GABA receptor
	glucoamylase	glucoamylase
10	glucoronidase	glucoronidase
	glycoprotein	glycoprotein
	Guanylyl	guanylylate cyclase
	helicase	helicase
	histone	histone
15	HOM	homologous
	homeobox	homeobox protein
	hydrolase	hydrolase
	hydroxysteroid	hydroxysteroid associated protein
	hypoxanthine	hypoxanthine associated protein
20	immunoglob	immunoglobulin
	immunoglobrecept	immunoglobulin receptor
	interferon	interferon
	interleukin	interleukin
	interleukinrecept	interleukin receptor
25	isomerase	isomerase
	isomeraseinhibitor	isomerase inhibitor
	isomerasereceptor	isomerase receptor
	kinase	kinase
	kinaseinhibitor	kinase inhibitor
30	kinasereceptor	kinase receptor
	kinesin	kinesin
	laminin	laminin associated protein
	lipase	lipase
	metallothionein	metallothionein
35	MHC	major histocompatibility complex
	misc_channel	miscellaneous channel
	ngf	nerve growth factor
	nuci_recpt	nuclear receptor
	nuclease	nuclease
40	oncogene	oncogene associated protein
	oxidase	oxidase
	oxygenase	oxygenase
	peptidase	peptidase
	peroxidase	peroxidase
45	phosphatase	phosphatase
	phosphataseinhib	phosphatase inhibitor

	phosphorylase PIR	phosphorylase PIR DATABASE (release 56, 29-OCT-1998)
5	polymerase potassium_channel prostaglandin protease proteaseinhib reductase	polymerase potassium channel protein prostaglandin protease protease inhibitor reductase
10	ribosomalprot RTR	ribosomal associated protein EMBLDATABASE translated entries not to be incorporated into SWISS-PROT (20-JUL-1998)
15	SIM SPTR	similar EMBL DATABASE translated entries to be incorporated into SWISS-PROT (20-JUL-1998)
20	struct sulfotransferase SWP	structural associated protein sulfotransferase SWISS-PROT DATABASE (release 18-OCT-1998)
25	SWPN synthase tgf tgfreceptor thioesterase thiolase tm7	SWISS-PROT Update (release 11-NOV-98) synthase transforming growth factor transforming growth factor receptor thioesterase thiolase
30	tnf traffic tnfreceptor TRN	seven transmembrane domain G-protein coupled receptor necrosis factor receptor tumor necrosis factor tumor trafficking associated protein EMBL DATABASE translated entries update (20-JUL-1998)
35	transcriptfactor transferase transport tubulin ubiquitin	transcription factor transferase transport protein tubulin ubiquitin
40	unclassified  water channel	Protein not categorized into one of the aforementioned protein families water channel protein

Column 5 of Table 1, entitled, "Cells or Tissues in Which Gene is Expressed", denotes tissues, represented by five digit numbers, in which RNA homologous to the ORF nucleic acid sequences is present. Tissues or cells corresponding to the numbers are provided in Table 2.

ORFX nucleic acids, and their encoded polypeptides, according to the invention are useful in a variety of applications and contexts. For example, various ORFX nucleic acids and polypeptides according to the invention are useful, *inter alia*, as novel members of the protein families indicated in Table 1, and/or according to the presence of domains and sequence relatedness to previously described proteins as summarized in Table 1.

ORFX nucleic acids and polypeptides according to the invention can also be used to identify cell types listed in Table 1 for an indicated ORFX according to the invention. Additional utilities for ORFX nucleic acids and polypeptides according to the invention are disclosed herein.

#### ORFX Nucleic Acids

The novel nucleic acids of the invention include those that encode an ORFX or ORFX-like protein, or biologically active portions thereof. The nucleic acids include nucleic acids encoding polypeptides that include the amino acid sequence of one or more of SEQ ID NO:2*n*, wherein *n* = 1 to 3161. The encoded polypeptides can thus include, *e.g.*, the amino acid sequences of SEQ ID NO: 2, 4, 6, 8, 10, . . . , 6310, 6312, 6314, 6316, 6318, 6320, and/or 6322.

In some embodiments, a nucleic acid encoding a polypeptide having the amino acid sequence of one or more of SEQ ID NO:2*n* (wherein *n* = 1 to 3161) includes the nucleic acid sequence of any of SEQ ID NO:2*n*-1 (wherein *n* = 1 to 3161), or a fragment thereof. Additionally, the invention includes mutant or variant nucleic acids of any of SEQ ID NO:2*n*-1 (wherein *n* = 1 to 3161), or a fragment thereof, any of whose bases may be changed from the disclosed sequence while still encoding a protein that maintains its ORFX-like activities and physiological functions. The invention further includes the complement of the nucleic acid sequence of any of SEQ ID NO:2*n*-1 (wherein *n* = 1 to 3161), including fragments, derivatives,

1781	95197259 (3581, 3582)	Novel Protein sim. GBank gij2114321[dbj BAA200371 - (D88733)] membrane glycoprotein [Equine herpesvirus 1]	Contains protein domain (PF00047) - glycoprotein Immunoglobulin domain	284488, 284686, 284687, 284768, 18108394, 264769, 18108397, 264259, 264691, 264692, 33657023, 264693, 264508, 264905, 264906, 264628, 264907, 264629, 264908, 264909, 264510, 265006, 264511, 265008, 264630, 265009, 264631, 264910, 264632, 264634, 264635, 264555, 264636, 264592, 264637, 264593, 264638, 18108381, 264639, 264758, 285010, 285011, 264602, 22279000, 264604, 264760, 264564, 264681, 264762, 264565, 264763, 264683, 264566, 264764, 264288, 264684, 264567, 18108354, 18108391, 264685, 264766
1782	87782690 (3583, 3584)	Novel Protein sim. GBank gij4337106[gb AAD18082] - (AF129756) BAT4 [Homo sapiens]	Contains protein domain (PF01585) - UNCLASSIFIED G-patch domain	22278997, 264259, 264508, 265007, 33657402, 87168559, 264369, 33657023, 35695855, 20281071, 264559, 18108387, 87168518
1783	95337877 (3585, 3586)	Novel Protein sim. GBank gij5579331[gb AAD45504.1]AF14573 - (AF145732) endoplasmic reticulum alpha-mannosidase I [Homo sapiens]	Contains protein domain (PF01532) - ATPase_associated Glycosyl hydrolase family 47	65274572, 22278995, 22278996, 22278997, 22278999, 264093, 264259, 29331824, 66714117, 60432289, 29331827, 29331828, 264103, 264105, 29331830, 265007, 264910, 265009, 60170831, 60433356, 21906754, 265010, 265017, 265019, 264681, 264682, 264288, 52844229, 21906765, 21906766, 21906767, 21906768, 21906769, 265020, 265021, 265022, 60170615, 52644150, 33657023, 33657109, 18108370, 18108374, 65274791, 20281071, 60432113, 22279000, 264482, 264584
1784	87759808 (3587, 3588)	Novel Protein sim. GBank gij4914604[emb CAB43577.1] - (AL050369) hypothetical protein [Homo sapiens]	Contains protein domain (PF01798) - UNCLASSIFIED Putative snoRNA binding domain	18108394, 22278995, 22278999, 264259, 29331822, 29331824, 29331825, 29146498, 29146499, 264508, 264905, 52644045, 264112, 265005, 265008, 264910, 60433356, 264757, 55812038, 87168474, 265011, 265017, 18108351, 264763, 264448, 264683, 264369, 21906765, 21906766, 21906767, 21906769, 29148784, 35695917, 60170615, 33657023, 264629, 18108374, 18108376, 35696423, 35695855, 264558, 264557, 264638, 264558, 18108385, 264564
1795	79747856 (3589, 3590)		UNCLASSIFIED	264632, 264635, 264636, 264595, 264596, 264907, 264566, 264909
1796	86599486 (3591, 3592)	Novel Protein sim. GBank gij585084[sp Q07803]IEFGM_RAT - ELONGATION FACTOR G, MITOCHONDRIAL PRECURSOR (MEF-G)	glycoprotein	264488, 264907, 264909, 264594, 264595, 264766, 264687, 21906765, 21906767, 264628, 264630, 264559

20							25					30				
Ser	Thr	Leu	Asp	Gly	Ala	Ala	Ala	Arg	Ala	Phe	Tyr	Glu	Ala	Leu	Ile	
		35					40					45				
Gly	Asp	Glu	Ser	Ser	Ala	Pro	Asp	Ser	Gln	Arg	Ser	Gln	Thr	Glu	Pro	
	50					55					60					
Ala	Arg	Glu	Arg	Lys	Arg	Lys	Arg	Arg	Ile	Met	Lys	Ala	Pro	Ala		
65					70				75					80		
Ala	Glu	Ala	Val	Ala	Glu	Gly	Ala	Ser	Gly	Arg	His	Gly	Gln	Gly	Arg	
				85					90					95		
Ser	Leu	Glu	Ala	Glu	Asp	Lys	Met	Thr	His	Arg	Ile	Leu	Arg	Ala	Ala	
			100				105					110				
Gln	Glu	Gly	Asp	Leu	Pro	Glu	Leu	Arg	Arg	Leu	Leu	Glu	Pro	His	Glu	
		115					120					125				
Ala	Gly	Gly	Ala	Gly	Gly	Asn	Ile	Asn	Ala	Arg	Asp	Ala	Phe	Trp	Trp	
	130					135					140					
Thr	Pro	Leu	Met	Cys	Ala	Ala	Arg	Ala	Gly	Gln	Gly	Ala	Ala	Val	Ser	
145					150					155				160		
Tyr	Leu	Leu	Gly	Arg	Gly	Ala	Ala	Trp	Val	Gly	Val	Cys	Glu	Leu	Ser	
				165					170					175		
Gly	Arg	Asp	Ala	Ala	Gln	Leu	Ala	Glu	Glu	Ala	Gly	Phe	Pro	Glu	Val	
			180					185				190				
Ala	Arg	Met	Val	Arg	Glu	Ser	His	Gly	Glu	Thr	Arg	Ser	Pro	Glu	Asn	
		195					200					205				
Arg	Ser	Pro	Thr	Pro	Ser	Leu	Gln	Tyr	Cys	Glu	Asn	Cys	Asp	Thr	His	
		210				215					220					
Phe	Gln	Asp	Ser	Asn	His	Arg	Thr	Ser	Thr	Ala	His	Leu	Leu	Ser	Leu	
225					230					235				240		
Ser	Gln	Gly	Pro	Gln	Pro	Pro	Asn	Leu	Pro	Leu	Gly	Val	Pro	Ile	Ser	
				245					250					255		
Ser	Pro	Gly	Phe	Lys	Leu	Leu	Leu	Arg	Gly	Gly	Trp	Glu	Pro	Gly	Met	
			260					265				270				
Gly	Leu	Gly	Pro	Arg	Gly	Glu	Gly	Arg	Ala	Asn	Pro	Ile	Pro	Thr	Val	
		275					280					285				
Leu	Lys	Arg	Asp	Gln	Glu	Gly	Leu	Gly	Tyr	Arg	Ser	Ala	Pro	Gln	Pro	
		290				295					300					
Arg	Val	Thr	His	Phe	Pro	Ala	Trp	Asp	Thr	Arg	Ala	Val	Ala	Gly	Arg	
305					310					315				320		
Glu	Arg	Pro	Pro	Arg	Val	Ala	Thr	Leu	Ser	Trp	Arg	Glu	Glu	Arg	Arg	
				325					330					335		
Arg	Glu	Glu	Lys	Asp	Arg	Ala	Trp	Glu	Arg	Asp	Leu	Arg	Thr	Tyr	Met	
			340					345				350				
Asn	Leu	Glu	Phe													
		355	</													

<210> 3585

<211> 2782

<212> DNA

<213> Homo sapiens

<400> 3585

60  
tcgcaggggca ggagaagcgg agctctcggg tcctctcagt cggacttcct gacgccgccca  
120

gtgggccccg ccccttgggc cgtcgccacc actgtagtca tgtacccacc gccgccgccg  
180  
ccgcctcacc gggacttcat ctcggtgacg ctgagctttg gcgagagcta tgacaacagc  
240  
aagagttggc ggcggcgctc gtgctggagg aaatggaagc aactgtcgag attgcagcgg  
300  
aatatgattc tcttctctct tgcctttctg cttttctgtg gactcctctt ctacatcaac  
360  
ttggctgacc attggaaagc tctggctttc aggctagagg aagagcagaa gatgaggcca  
420  
gaaattgctg ggttaaaacc agcaaatacca cccgtcttac cagctcctca gaaggcggac  
480  
accgacctg agaacttacc tgagatttct tcacagaaga cacaagaca catccagcgg  
540  
ggaccacctc acctgcagat tagaccccc aaccaagacc tgaaggatgg gaccaggag  
600  
gaggccacaa aaaggcaaga agccccgtg gatccccgcc cggaaggaga tccgcagagg  
660  
acagtcatca gctggagggg agcgggtgatc gagcctgagc agggcaccga gctcccttca  
720  
agaagagcag aagtggccac caagcctccc ctgccaccgg ccaggacaca gggcacacca  
780  
tgcatctga actatcgcca gaaggcgctg attgacgtct tctgcatgc atggaaagga  
840  
taccgcaagt ttgcatgggg ccatgacgag ctgaagcctg tgtccaggtc cttcagtgg  
900  
tggtttggcc tgggtctcac actgatcgac gcgctggaca ccatgtggat cttgggtctg  
960  
aggaaagaat ttgaggaagc caggaaagtg gtgtcgaaga agttacactt tgaaggagc  
1020  
gtggacgtca acctgtttga gagcacgac cgcctcctgg gggggctcct gaggcctac  
1080  
cacctgtctg gggacagcct cttcctgagg aaagctgagg attttgaaa tcggctaattg  
1140  
cctgccttca gaacaccatc caagattcct tactcggatg tgaacatcgg tactggagtt  
1200  
gcccacccgc caccgtggac ctccgacagc actgtggccg aggtgaccag cattcagctg  
1260  
gagttccggg agctctcccg tctcacaggg gataagaagt ttcaggaggc agtggagaag  
1320  
gtgacacagc acatccacgg cctgtctggg aagaaggatg ggctggtgcc catgttcatc  
1380  
aataccacaa gtggcctctt caccacctg ggcgtattca cgctgggcgc cagggccgac  
1440  
agctactatg agtacctgct gaagcagtg atccaggggc ggaagcagga gacacagctg  
1500  
ctggaagact acgtggaagc catcgagggt gtcagaacgc acctgctgcg gcactccgag  
1560  
cccagtaagc tcacctttgt gggggagctt gccacggcc gcttcagtgc caagatggac  
1620  
cacctgggtg gcttctgcc agggacgctg gctctgggcg tctaccacgg cctgcccgc  
1680  
agccacatgg agctggccca ggagctcatg gagacttggt accagatgaa ccggcagatg  
1740

gagacggggc tgagtcccca gatcgtgcac ttcaaccttt acccccagcc gggccgctcg  
 1800  
 gacgtggagg tcaagccagc agacaggcac aacctgctgc ggccagagac cgtggagagc  
 1860  
 ctgtttctacc tgtaccgcgt cacaggggac cgcaaatacc aggactgggg ctgggagatt  
 1920  
 ctgcagagct tcagccgatt cacacgggtc ccctcgggtg gctattcttc catcaacaat  
 1980  
 gtccaggatc ctgagaagcc cgagcctagg gacaagatgg agagcttctt cctgggggag  
 2040  
 acgctcaagt atctgttctt gctcttctcc gatgacccaa acctgctcag cctggacgcc  
 2100  
 tacgtgttca acaccgaagc ccaccctctg cctatctgga cccctgccta ggggtggatgg  
 2160  
 ctgctgggtg ggggacttcg ggtgggcaga ggcaccttgc tgggtctgtg gcattttcca  
 2220  
 agggcccacg tagcaccggc aaccgccaag tggcccaggc tctgaactgg ctctgggctc  
 2280  
 ctctcgtct ctgctttaat caggacaccg tgaggacaag tgaggccgct agtcttggtg  
 2340  
 tgatgccccg tgggctgggc cgctggagcc tccgcctgct tcctccagaa gacacgaatc  
 2400  
 atgactcacg attgctgaag cctgagcagg tctctgtggg ccgaccagag gggggcttcg  
 2460  
 aggtggtccc tgggtactgg gtgaccgagt ggacagccca ggggtgcagct ctgccccggc  
 2520  
 tcgtgaagcc tcagatgtcc ccaatccaag ggtctggagg ggctgccgtg actccagagg  
 2580  
 cctgaggctc cagggtgggc tctggtgttt acaagctgga ctcagggatc ctctgggccc  
 2640  
 ccccgagggg ggcttgagg gctggacggc aagtccgtct agctcacggg cccctccagt  
 2700  
 ggaatgggtc ttttcggtgg agataaaaagt tgatttgctc taaaaaaaaa aaaaaaaaaa  
 2760  
 aaaaaaaaaa aaaaaaaaaa aa  
 2782

&lt;210&gt; 3586

&lt;211&gt; 663

&lt;212&gt; PRT

&lt;213&gt; Homo sapiens

&lt;400&gt; 3586

Met	Tyr	Pro	Pro	Pro	Pro	Pro	Pro	His	Arg	Asp	Phe	Ile	Ser	Val
1				5				10					15	
Thr	Leu	Ser	Phe	Gly	Glu	Ser	Tyr	Asp	Asn	Ser	Lys	Ser	Trp	Arg
			20					25				30		
Arg	Ser	Cys	Trp	Arg	Lys	Trp	Lys	Gln	Leu	Ser	Arg	Leu	Gln	Arg
		35				40					45			
Met	Ile	Leu	Phe	Leu	Leu	Ala	Phe	Leu	Leu	Phe	Cys	Gly	Leu	Phe
	50				55					60				
Tyr	Ile	Asn	Leu	Ala	Asp	His	Trp	Lys	Ala	Leu	Ala	Phe	Arg	Leu
65				70					75				80	
Glu	Glu	Gln	Lys	Met	Arg	Pro	Glu	Ile	Ala	Gly	Leu	Lys	Pro	Ala

				85					90					95		
Pro	Pro	Val	Leu	Pro	Ala	Pro	Gln	Lys	Ala	Asp	Thr	Asp	Pro	Glu	Asn	
			100					105					110			
Leu	Pro	Glu	Ile	Ser	Ser	Gln	Lys	Thr	Gln	Arg	His	Ile	Gln	Arg	Gly	
		115					120					125				
Pro	Pro	His	Leu	Gln	Ile	Arg	Pro	Pro	Ser	Gln	Asp	Leu	Lys	Asp	Gly	
		130				135					140					
Thr	Gln	Glu	Glu	Ala	Thr	Lys	Arg	Gln	Glu	Ala	Pro	Val	Asp	Pro	Arg	
145					150					155					160	
Pro	Glu	Gly	Asp	Pro	Gln	Arg	Thr	Val	Ile	Ser	Trp	Arg	Gly	Ala	Val	
				165					170					175		
Ile	Glu	Pro	Glu	Gln	Gly	Thr	Glu	Leu	Pro	Ser	Arg	Arg	Ala	Glu	Val	
			180					185					190			
Pro	Thr	Lys	Pro	Pro	Leu	Pro	Pro	Ala	Arg	Thr	Gln	Gly	Thr	Pro	Val	
		195					200					205				
His	Leu	Asn	Tyr	Arg	Gln	Lys	Gly	Val	Ile	Asp	Val	Phe	Leu	His	Ala	
		210				215					220					
Trp	Lys	Gly	Tyr	Arg	Lys	Phe	Ala	Trp	Gly	His	Asp	Glu	Leu	Lys	Pro	
225					230					235					240	
Val	Ser	Arg	Ser	Phe	Ser	Glu	Trp	Phe	Gly	Leu	Gly	Leu	Thr	Leu	Ile	
				245					250					255		
Asp	Ala	Leu	Asp	Thr	Met	Trp	Ile	Leu	Gly	Leu	Arg	Lys	Glu	Phe	Glu	
			260					265					270			
Glu	Ala	Arg	Lys	Trp	Val	Ser	Lys	Lys	Leu	His	Phe	Glu	Lys	Asp	Val	
			275				280					285				
Asp	Val	Asn	Leu	Phe	Glu	Ser	Thr	Ile	Arg	Ile	Leu	Gly	Gly	Leu	Leu	
		290				295					300					
Ser	Ala	Tyr	His	Leu	Ser	Gly	Asp	Ser	Leu	Phe	Leu	Arg	Lys	Ala	Glu	
305					310					315					320	
Asp	Phe	Gly	Asn	Arg	Leu	Met	Pro	Ala	Phe	Arg	Thr	Pro	Ser	Lys	Ile	
			325						330					335		
Pro	Tyr	Ser	Asp	Val	Asn	Ile	Gly	Thr	Gly	Val	Ala	His	Pro	Pro	Arg	
			340					345					350			
Trp	Thr	Ser	Asp	Ser	Thr	Val	Ala	Glu	Val	Thr	Ser	Ile	Gln	Leu	Glu	
		355					360					365				
Phe	Arg	Glu	Leu	Ser	Arg	Leu	Thr	Gly	Asp	Lys	Lys	Phe	Gln	Glu	Ala	
		370				375					380					
Val	Glu	Lys	Val	Thr	Gln	His	Ile	His	Gly	Leu	Ser	Gly	Lys	Lys	Asp	
385					390					395					400	
Gly	Leu	Val	Pro	Met	Phe	Ile	Asn	Thr	His	Ser	Gly	Leu	Phe	Thr	His	
				405					410					415		
Leu	Gly	Val	Phe	Thr	Leu	Gly	Ala	Arg	Ala	Asp	Ser	Tyr	Tyr	Glu	Tyr	
			420					425								

515	520	525
Thr Gly Leu Ser Pro Glu Ile Val His Phe Asn Leu Tyr Pro Gln Pro		
530	535	540
Gly Arg Arg Asp Val Glu Val Lys Pro Ala Asp Arg His Asn Leu Leu		
545	550	555
Arg Pro Glu Thr Val Glu Ser Leu Phe Tyr Leu Tyr Arg Val Thr Gly		
565	570	575
Asp Arg Lys Tyr Gln Asp Trp Gly Trp Glu Ile Leu Gln Ser Phe Ser		
580	585	590
Arg Phe Thr Arg Val Pro Ser Gly Gly Tyr Ser Ser Ile Asn Asn Val		
595	600	605
Gln Asp Pro Gln Lys Pro Glu Pro Arg Asp Lys Met Glu Ser Phe Phe		
610	615	620
Leu Gly Glu Thr Leu Lys Tyr Leu Phe Leu Leu Phe Ser Asp Asp Pro		
625	630	635
Asn Leu Leu Ser Leu Asp Ala Tyr Val Phe Asn Thr Glu Ala His Pro		
645	650	655
Leu Pro Ile Trp Thr Pro Ala		
660		

&lt;210&gt; 3587

&lt;211&gt; 3148

&lt;212&gt; DNA

&lt;213&gt; Homo sapiens

&lt;400&gt; 3587

```

nctttttttt ttttttttga gtgtggggtc agtttattgg gcatgcgtca gtcagaggct
60
gggctggcca gggtcgggta gggcagcagt ttgtctggac cccgagaaac ccaactggaa
120
tccagggcct catctgcttc aaagccaaag tcttctctcaa ccttaatctg caccggggcc
180
agctctggag tcagcgcatt tcctgctcgg cgtccatccc gtggcactcg ccgcctcttc
240
cgcccaactgg gccctcacc gggggctggg ctgccgggtt ctgggggtgc aggagtcctt
300
ctggggcgggg acagtgtctc tttctctgga ggctcattct ccgcattgcc tgggggtgggg
360
gcatccgtgc cctggctgcc ctcatcctcc agcacaatgg tgaactggct ggcccggtag
420
tcatccccgt aggagtccag cactctcatg aggaacctcc gttcctgctg cagcctccga
480
gttatectct gcacctgatg gagcctgttc aggaccgct cgttcacctg ctcgatctcc
540
cggcagcgcc gacctagtgc ctggtacttt ctgcgattta attcccgtg gcgcccgcgc
600
cgaccccggg ctgcctcttc ctcttcctct cgctcccgga gccctgagcc gccagacca
660
cctgacacaa actccacttc cgtctccagc tcgctctcca ggatgtggcc accaaatagg
720
ggaggcaacg ccaactctga gcctggcggc gctgagaact cctcaaagcc cacggctgcc
780
atgggtccctt ctctctgctc caattccatc tccgcgacct ccggaagccc cgggcctcag
840

```